Applied and Environmental Microbiology, Dec. 2011, p. 8795–8798 Vol. 77, No. 24
0099-2240/11/$12.00 doi:10.1128/AEM.05491-11

Evaluation of Subsampling-Based Normalization Strategies for Tagged High-Throughput Sequencing Data Sets from Gut Microbiomes▽

Daniel Aguirre de Cárcer, Stuart E. Denman, Chris McSweeney, and Mark Morrison*

CSIRO Preventative Health National Research Flagship and Division of Livestock Industries, Queensland Bioscience Precinct, St. Lucia, QLD 4067, Australia

Received 16 May 2011/Accepted 27 September 2011

Several subsampling-based normalization strategies were applied to different high-throughput sequencing data sets originating from human and murine gut environments. Their effects on the data sets' characteristics and normalization efficiencies, as measured by several β-diversity metrics, were compared. For both data sets, subsampling to the median rather than the minimum number appeared to improve the analysis.

The high-throughput sequencing of tagged hypervariable regions of the bacterial 16S rRNA genes is rapidly becoming one of the methods of choice in the analysis of complex microbial communities (10). Due to technical reasons (e.g., imperfect pooling prior to sequencing and/or stochastic events during sequencing [4]) or when comparing samples from different sequencing rounds, the amounts of sequences obtained per sample/tag differ. Furthermore, as the coverage for a given sample increases, sequences are added arithmetically, but the number of operational taxonomic units (OTUs) increases at a decreasing, logarithmic pace. For these reasons, a normalization step prior to analysis is widely used to standardize sampling efforts and bring the data from different samples onto a common scale. One way to handle this issue is to randomly subsample each community to a common depth, a procedure often referred to as rarefaction. Rarefaction approaches have commonly been used in ecology to evaluate sampling effort and community richness (1, 3) and more recently with β-diversity measures (5). Rarefaction is also included as a normalization step in widely used microbial community analysis pipelines, such as QIIME (2).

Recent papers using subsampling as a normalization step prior to β-diversity analysis have reported subsampling to the lowest number of sequences produced from any sample (5) or even less (7), while others appear to have used arbitrarily defined thresholds (6). The rationale behind subsampling depth choice is generally unreported but presumably strives to strike a compromise between information loss and data set balance. Even though decreasing the subsampling depth can improve a data set's balance, it could also lead to the suboptimal use of the information contained in the data set. The trade-off between number of samples and depth of coverage, together with the performance of different analytical techniques, has recently been explored (5a). However, to the best of our knowledge, there does not appear to have been a systematic evaluation of the relationships between the depth of subsampling as a normalization strategy with information loss, data set balance, and efficacy for the analysis of tagged high-throughput sequencing data sets. We describe here our comparison of the normalization efficiency of different subsampling depths and a recodification strategy (recoding singletons as zeros) on the β-diversity measures produced from two different data sets derived from gut microbiomes.

Two different data sets generated in our laboratory were used for these studies. One was produced from human colon mucosa biopsy specimens (data set Q; 40 samples, 455,660 sequences), and the second was produced from mice cecal mucosa and fecal samples (data set D; 46 samples, 194,663 sequences). Both sample sets were processed independently. Each sample was tagged and amplified in triplicate (thus three different technical replicates, each with a different tag, were generated for each sample) using primers flanking the V1 to V3 regions of the bacterial *rrs* gene. Equimolar amounts of each replicate were pooled and sequenced using a Roche 454 FLX sequencer with titanium chemistry. The sequences obtained were processed using QIIME; sequences were assigned to samples using the tag information, filtered for correct length and quality thresholds, and grouped in OTUs at a 0.97 distance threshold. Those OTUs not appearing in at least two replicates across the data set were discarded to eliminate noise and possible artifacts. Some tags failed to provide a significant number of sequences and were eliminated from further analysis.

The resulting data sets were normalized using the following strategies: (i) rarefaction (Rare), randomly subsampling each sample to a common depth; (ii) rarefaction and recodification (Rare+Recode), the same as Rare but deletes singletons (i.e., recoding 1 as zero); (iii) multiple rarefaction (MultiRare), which randomly subsamples each sample to a common depth 100 times and then uses the average; and (iv) multiple rarefaction and recodification (MultiRare+Recode), which is the same as MultiRare but recodes values lower than 1.01 as zero. The subsampling depths employed were based on the data sets' characteristics (quartiles): (i) initial, no subsampling; (ii) 75%, subsampling to the higher quartile; (iii) 50%, subsampling to the median; (iv) 25%, subsampling to the lower quartile; and (v) Min, subsampling to the smallest coverage in the data set.

* Corresponding author. Mailing address: Queensland Bioscience Precinct, 306 Carmody Road, St. Lucia, QLD 4068, Australia. Phone: 61 7 3214 2216. Fax: 61 7 3214 2900. E-mail: mark.morrison@csiro.au.

▽ Published ahead of print on 7 October 2011.

TABLE 1. Effect of the different strategies on the distributions of the data sets[a]

Strategy	Data set Q			Data set D		
	No. of OTUs	No. of sequences/ replicate ± SD	Total no. of sequences	No. of OTUs	No. of sequences/ replicate ± SD	Total no. of sequences
Total	3,091	3,997 ± 2,116	455,660	17,015	1,421 ± 626	194,663
Initial	2,042	3,981 ± 2,103	453,941	7,120	1,327 ± 579	181,860
Rare 75%	2,037	3,514 ± 1,225	400,592	7,115	1,243 ± 439	170,255
MultiRare1 75%	2,042	3,514 ± 1,224	400,592	7,120	1,243 ± 439	170,255
Rare+Recode1 75%	1,628	3,421 ± 1,218	390,005	4,079	954 ± 361	130,696
MultiRare+Recode1 75%	1,669	3,426 ± 1,225	390,587	4,228	960 ± 369	131,587
Rare+Recode2 75%	1,628	3,421 + 1,218	390,005	3,913	1,123 + 476	153,981
Rare+Recode5 75%	928	3,218 + 1,195	366,900	885	742 + 379	101,708
Rare+Recode10 75%	560	2,999 + 1,163	341,969	293	511 + 310	70,080
Rare 50%	2,020	2,755 ± 430	314,057	7,040	1,057 ± 268	144,853
MultiRare1 50%	2,042	2,755 ± 430	314,057	7,120	1,057 ± 268	144,853
Rare+Recode1 50%	1,520	2,665 ± 427	303,821	3,635	789 ± 212	108,048
MultiRare+Recode1 50%	1,591	2,673 ± 432	304,773	4,188	812 ± 228	111,281
Rare+Recode2 50%	1,520	2,665 + 426	303,821	3,656	789 + 212	108,152
Rare+Recode5 50%	809	2,474 + 422	282,094	705	490 + 160	67,236
Rare+Recode10 50%	461	2,280 + 430	259,993	236	310 + 129	42,540
Rare 25%	2,004	2,477 ± 293	282,345	6,642	744 ± 78	101,983
MultiRare1 25%	2,042	2,477 ± 293	282,345	7,120	744 ± 78	101,983
Rare+Recode1 25%	1,456	2,388 ± 289	272,263	2,763	520 ± 70	71,294
MultiRare+Recode1 25%	1,530	2,397 ± 292	273,289	3,377	537 ± 79	73,637
Rare+Recode2 25%	1,456	2,388 + 289	272,263	2,763	520 + 70	71,294
Rare+Recode5 25%	757	2,206 + 289	251,577	464	295 + 65	40,483
Rare+Recode10 25%	421	2,021 + 304	230,441	174	163 + 59	23,059
Rare Min	1,476	450 ± 0	51,300	5,696	435 ± 0	59,595
MultiRare1 Min	2,042	451 ± 0	51,414	7,120	439 ± 0	60,139
Rare+Recode1 Min	712	397 ± 19	45,220	1,748	269 ± 31	36,873
MultiRare+Recode Min	639	393 ± 21	44,779	1,918	272 ± 42	37,331
Rare+Recode2 Min	712	396 + 19	45,220	1,747	269 + 31	36,873
Rare+Recode5 Min	245	323 + 40	36,912	267	199 + 32	17,771
Rare+Recode10 Min	112	265 + 51	30,258	103	60 + 29	8,295

[a] In the case of the MultiRare strategies, sequence values are based on averages of the iterations. SD, standard deviation.

In all cases, replicates possessing fewer sequences than the subsampling threshold were kept as they were.

It could be expected that a comparison of sequences obtained from replicates taken from a given sample would show similarity (not accounting for technical noise). Therefore, the most effective normalization approach would be the one that exhibits the least distance between replicates from the same sample. For each combination and strategy of subsampling depth, we generated between-replicate distance (or other metrics) matrices using the Euclidean, Bray-Curtis, unweighted UniFrac (8), and Rao diversity (9) measures using several R packages (Vegan, Ade4, Picante). Then, for each replicate, we obtained a resolution value: the ratio (Rt) of average distance (or other metric) to the replicates from the same sample (average within-sample distance) divided by the average distance to all the replicates in the data set (average distance to all replicates). Next, the average of the Rt values obtained for each particular strategy combination and subsampling depth was adopted as a proxy of its resolution (Rt = average within-sample distance/average distance to all replicates; the lower the number, the greater the resolution), and the results were plotted. The observed differences between selected strategies were statistically tested by comparing the Rt values for each replicate on a paired two-sided Wilcox test (to maintain within-sample independence, one replicate per sample was removed from the analysis; $\alpha = 0.05$, $n = 91$ and 74 samples for data sets D and Q, respectively) using R packages.

Table 1 shows that the different normalization approaches had similar effects on the distributions of both data sets; average sequences per replicate (or total number of sequences) and their standard deviations serve as proxies for total information and balance of the data sets, respectively. There was a constant decrease in total sequences and sequences per replicate with decreasing subsampling depth (Table 1), with a concomitant decrease of the standard deviation (increased balance) until the minimum depth, when the decrease was sharpest. In comparison to subsampling to the same depth, the addition of the recodification strategy did not exhibit much of an effect on the amount of sequences per sample but strongly reduced the number of OTUs in the data sets. It was also observed that elevating the recodification threshold translated into an increasingly greater loss of data (unpublished data). The multiple rarefaction strategies behaved in a manner similar to that of the single rarefaction strategies, except that the former did not decrease the number of total OTUs with decreasing subsampling depth. This was because the samples' values for such OTUs where abundances were below 1.01 were retained, since such an effect disappears in the MultiRare+Recode approach.

The effects of the different normalization approaches employed on resolution were concordant in both data sets (Fig. 1). The results based on Euclidean distances showed no effects due to the recodifications applied, and there was a trend of increased resolution with decreasing depth up to the 50%

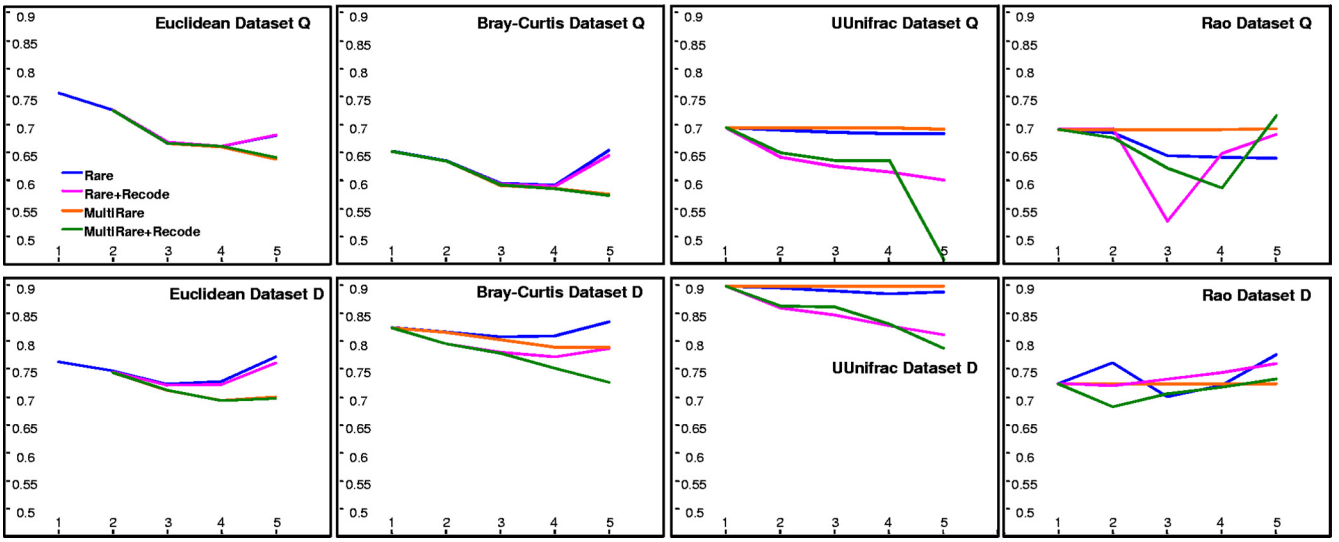

FIG. 1. Effect of the different normalization strategies (color lines) applied on resolution. *y* axis, average Rt value (the lower the number, the greater the resolution); *x* axis, rarefaction depth (1, initial; 2, 75%; 3, 50%; 4, 25%; 5, Min). Panel label represents metric used and data set origin. Plots are drawn as line graphs, and *y* axes are focused to help emphasize the trends across different subsampling depths and strategies.

depth in the case of single rarefaction or even lower in the case of the MultiRare strategies. The Bray-Curtis dissimilarity results showed a similar behavior, except that for data set D (having many more OTUs but fewer sequences per replicate), the recodification strategies seem to have improved the resolution. The results using the unweighted UniFrac metric show an increased resolution related to the depth of subsampling only for the recodification strategies. Regarding the Rao diversity, no changes were observed for the multiple rarefaction approach. Both recodification strategies seemed to improve resolution initially and then drastically reduced it, although the coverage quartile at which the phenomenon occurred was different for the two data sets. Normal rarefaction improved resolution when subsampling to the median in both data sets, but no further improvements were observed with further subsampling depth.

The differences observed using the different metrics arise from their different characteristics: the Euclidean distance is more affected by extreme values but relatively insensitive to small changes in absolute abundance, hence the observed null effect of recoding on the resolution. The Bray-Curtis dissimilarity does not suffer from the double zero problem and gives equal weight to all species and samples, which might explain the increased resolution caused by the recodification strategies compared to that observed for the Euclidean distance. The unweighted UniFrac and Rao diversity measures take into account the phylogenetic information of each OTU; it is therefore the overall phylogenetic resemblance between samples that matters. However, the unweighted UniFrac takes into account only presence/absence data, explaining the null effect of subsampling if not accompanied by a recodification step. The Rao diversity is a weighted measure and thus benefited from some degree of subsampling.

In summation, our results suggest that subsampling to the minimum as a normalization strategy did not perform particularly well, with data sets presenting some degree of coverage heterogeneity. On the other hand, subsampling to the median in all cases either improved the analysis or had no effect but still retained a larger proportion of the initial sequences. It also seems that the recodification strategy was worth applying, because it did not reduce the resolution of the analysis and in several instances improved it. In this sense, the MultiRare+Recode 50% strategy in most cases substantially ($P < 0.05$) improved the resolution of the analyses of both data sets compared to both the initial and RareMin strategies, using the four metrics. The exceptions were limited to Rao diversity measurements, where only the MultiRare+Recode 50% strategy was significantly different from the RareMin strategy in data set D. For these reasons, the subsampling strategy should be carefully considered and described when analyzing data sets comprised of samples produced from different sequencing runs and/or that have significant differences in sample coverage.

This research has been supported with funds provided by a CSIRO OCE Science Leader award (to M.M.) and funds from CSIRO's Transformational Biology Capability Platform.

We thank Antonio Reverter-Gomez and David Lovell for critical reading of the manuscript. We are also grateful to Barbara Leggett (Queensland Institute of Medical Research) and Ranjeny Thomas (Princess Alexandra Hospital, Brisbane) for our extended use of the data sets arising from our collaborations, as well as Rob Moore and Honglei Chen for the amplicon library preparation and sequencing.

All authors conceived the experiment. D.A.D.C., S.E.D., and M.M. cowrote the paper. D.A.D.C. designed the experiment and carried out the data analysis.

REFERENCES

1. **Brewer, A., and M. Williamson.** 1994. A new relationship for rarefaction. Biodivers. Conserv. **3:**373–379.
2. **Caporaso, J. G., et al.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Meth. **7:**335–336.
3. **Gotelli, N. J., and R. K. Colwell.** 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. **4:**379–391.
4. **Harris, J. K., et al.** 2010. Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. Appl. Environ. Microbiol. **76:**3863–3868.

5. **Horner-Devine, M. C., M. Lage, J. B. Hughes, and B. J. M. Bohannan.** 2004. A taxa-area relationship for bacteria. Nature **432:**750–753.
5a. **Kuczynski, J., C. Lozupone, N. Fierer, and R. Knight.** 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat. Methods **7:**813–819.
6. **Lauber, C. L., M. Hamady, R. Knight, and N. Fierer.** 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. **75:**5111–5120.
7. **Lauber, C. L., N. Zhou, J. I. Gordon, R. Knight, and N. Fierer.** 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol. Lett. doi:10.1111/j.1574-6968.2010.01965.x.
8. **Lozupone, C., and R. Knight.** 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. **71:**8228–8235.
9. **Rao, C. R.** 1982. Diversity and dissimilarity coefficients—a unified approach. Theor. Popul. Biol. **21:**24–43.
10. **Roh, S. W., G. C. J. Abell, K.-H. Kim, Y.-D. Nam, and J.-W. Bae.** 2010. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends Biotechnol. **28:**291–299.